



Automatic identification method for driving risk status based on multi-sensor data

Lixin Yan¹ · Yike Gong¹ · Zhijun Chen^{2,3} · Zhenyun Li¹ · Junhua Guo¹

Received: 11 January 2021 / Accepted: 28 May 2021 / Published online: 14 September 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract

Real risk status detection is an effective way to reflect risky or dangerous driving behaviors and therefore to prevent road traffic accidents. However, a driver's risk status is not only difficult to define but also uncontrollable and uncertain. In this study, a simulated experiment with 30 drivers was conducted using a driving simulator to collect the multi-sensor data of road conditions, humans, and vehicles. The driving risk status was classified into three states (0 - incident, 1 - near crash, or 2 - crash) on the basis of the playback system of the driving simulator. The experimental data were pre-processed using the cubic spline interpolation method and the time-windows theory. A driving risk status identification model was established using the C5.0 decision tree algorithm, and the receiver operating characteristic curve (ROC) was adopted to evaluate the performance of the identification model. The results indicated that respiration (RESP), vehicle speed (SPE), SM_FATIGUE, distance to the left lane (LLD), course angle (CA), and skin conductivity (SC) had a significant correlation ($p < 0.05$) with the driving risk status. The identification accuracy of the C5.0 decision tree algorithm was 78%, and the areas under the ROC were 0.934, 0.77, and 0.845, respectively. Moreover, compared with other four identification algorithms, the algorithm performance evaluation indexes TPR (0.780), precision (0.753), recall (0.78), F-measure (0.756), and kappa (0.884) of the C5.0 decision tree were all the best. The conclusion can provide reference evidence for danger warning systems and intelligent vehicle design.

Keywords C5.0 decision tree · Identification algorithm · Multi-sensors · Risk status · Traffic safety

1 Introduction

Road safety is an important issue in the transport field. It is determined by drivers, vehicles, and the driving environment. Previous research has revealed that more than 90% of the traffic accidents are associated with unsafe driving behaviors [1]. Real risk status detection is an effective way to reflect risky or dangerous driving behaviors and therefore to prevent road traffic accidents. However, the drivers' risk status is not only difficult to define but also uncontrollable and uncertain.

Dingus et al. [2] defined three types of traffic events as follows:

Crash: situations in which there is physical contact between the subject vehicle and another vehicle, fixed object, pedestrian, cyclist, or animal.

Near crash: situations requiring a rapid, severe, and evasive maneuver to avoid a crash.

Incident: situations requiring an evasive maneuver of less magnitude than for a near crash.

The different risk statuses of the drivers will result in the three different abovementioned traffic events. If an automatic alarm system or an automatic driving system can identify the driver's current driving risk status, it can avoid accidents or, in the case of a dangerous status, remind or take over the driver and complete the driving task safely [3–5].

The purpose of the present study was to classify and predict the drivers' risk status through a large amount of data from a driving simulation experiment based on an advanced decision tree (C5.0) algorithm. The decision tree visually explains the

✉ Zhijun Chen
chenzj556@whut.edu.cn

¹ School of Transportation and Logistics, East China Jiaotong University, Nanchang, China

² Intelligent Transport Systems Research Center, Wuhan University of Technology, Wuhan, China

³ Engineering Research Center for Transportation Safety, Ministry of Education, Wuhan, China

relationship between the risk status and the related factors. In addition, the decision tree can achieve high prediction accuracy. Therefore, in this paper, we propose a C5.0 algorithm based on the strong correlation between multiple information and driving risk to predict the drivers' risk status.

For real risk status detection, Charlton [6] investigated the acceleration variation caused by the attention lapses in the curves. Caird et al. [7] found that drivers' attention lapses would emerge with a decrease in the driving speed. Guo [8] pointed out that the effective identification and early warning for a dangerous status of driving behaviors are crucial for preventing road traffic accidents. The basic thought of the thesis is to use various cluster methods to classify the drivers' status variation in the statistical dimension. The adopted state variables included driving behavioral feature parameters and vehicle state parameters.

A classification of driving styles was used as a surrogate to reveal the potential driving risks in the existing research. Acceleration and braking, time headway control, lane change, and turning control were the main indexes for judgment [9, 10]. Wang et al. [11] used a vehicle's longitudinal acceleration and acceleration jerk as the evaluation indexes and developed a driving style detection system, which yielded a satisfactory result in terms of improving the driving behavior and riding comfort. Simons-Morton et al. [12] found that the drivers overestimated driving abilities and had a relatively huge bias for estimating the time headway. An incorrect estimation poses a considerable threat for road safety. Macadam [13] adopted the naturalistic driving data of 36 drivers to investigate the relationship between the driving style and the car-following behavior. The results showed that the frequency of a close car-following phenomenon is strongly related with the driving style.

Except for the direct parameters that can reveal the real risk status, several indirect indexes need to be adopted to reflect dangerous driving behaviors, such as distracted driving and fatigue driving. There have been a number of studies showing a driver's eye movement characteristics related to the above unsafe driving behavior. Dehban et al. [14] used a cognitive-based driver's steering behavior modeling method to explain how the driver can acquire information in his/her visual field and how the driver manipulates his/her environment. Hills et al. [15] explored the vertical eye movement carryover from one task to another task and found that it is a potentially distracting effect on the safety of novice drivers. Lantieri et al. [16] explored the effect of gateways to reduce the amount of distraction, by analyzing the drivers' eye movement data. Li et al. [17] studied the visual scanning behavior of drivers at signalized and unsignalized intersections, and found that intersection types made differences on drivers' scanning behavior. As for a fatigue driving-related study, Jimenez-Pinto et al. [18] obtained the shape of the eyes and the mouth to predict whether the driver was yawning or

blinking. Zhang et al. [19] carried out a preliminary estimation of the eye gaze from the elliptical features of an iris and obtained the vectors describing the translation and the rotation of the eyeball.

From the perspective of methodology, machine learning algorithms have been applied as a typical data analysis method in road safety studies. Abdelwahab et al. [20] applied an artificial neural network model to predict the traffic accident risk at signalized intersections. The results showed that the multi-layer perceptron (MLP) neural network model can achieve ideal accuracy in intersection risk prediction, which indicates that the modified algorithm has good generalization performance. Hernandezgress et al. [21] combined multi-sensory information and used a principal component analysis and neural networks to identify whether a driver was behaving normally. One of the main problems of machine learning is the uncertainty of extending a given model to a new problem. To determine whether an algorithm can be generalized well, the datasets are divided into two (training and testing) or three (training, validation, and testing) datasets for validation. Huang et al. [22], using convolutional neural network (CNN) for visual analysis, proposed a hybrid CNN framework (HCF) to detect the behavior of distracted drivers, and deep learning is used to process image features to help drivers maintain safe driving habits. Jeberson et al. [23, 24] used fog computing with the Internet of Things and machine learning to achieve intelligent healthcare data segregation.

Moreover, Wang et al. [25] used classified regression trees to reveal the relationship between the driving risk level and the influencing factors from three aspects of the road environment, driver characteristics, and vehicle characteristics. Li et al. [26] proposed a feasible method of data analysis, learning, and parameter calibration based on an RBF neural network to determine the corresponding decision support system on the basis of the fact that an integrated simulation platform for urban traffic was built. The Bayesian network (BN) classification algorithm has flexibility in the classification of accident datasets. It uses the previous data information and increases the experience of decision makers. AIKheder et al. [27] on the basis of three models comprehensively analyzed the correlative factors of traffic accident severity; the results showed that BNs are more accurate in predicting multiple variables than other algorithms. Cura et al. [28] used LSTM and CNN's neural network model to classify and evaluate bus driver behavior characterized by deceleration, engine acceleration pedals, turns, and lane change attempts. Kumagai et al. [29] proposed a dynamic BN algorithm to predict a dangerous driving behavior, and the vehicle's speed and acceleration were regarded as the related variables. In addition, other machine learning algorithms such as SVM [30] were also applied as prediction algorithms for risk status identification.

As there are many factors that affect the driving risk, various methods should be combined to improve the accuracy of

driving risk classification and prediction. Yan et al. [31] used BNs to analyze the main factors that significantly affect the driving risk status and selected the most relevant factors to establish the driving risk status prediction model logistically. Panagopoulos et al. [32] used an extreme gradient boosting (XGB) algorithm that provides a short-term forecast for dangerous driving behaviors.

Previous studies have focused on exploring the relationships among driving risk status, driver personality characteristics, vehicle characteristics, road conditions, and the driving environment. Although there are part of the study focused on the use of sensors to assess the risks of driving, not many types of data were collected by the sensors. Therefore, in this paper, we propose a C5.0 algorithm based on the strong correlation between multi-information and the driving risk. Then, a driving risk state detection model based on the relevant factors is established by using the C5.0 algorithm. Finally, the ROC curve is applied to judge the classification and detection results. Then, the true positive rate (TPR), false positive rate (FPR), precision (P), recall rate (R), F-measure, and kappa value were applied to evaluate the performance of different prediction models.

The remainder of this paper is organized as follows. Section 2 introduces the data processing method adopted in this study and describes in detail the working principle of the C5.0 decision tree. Section 3 describes the RS experiment protocol and data collection. Sections 4 and 5 discuss the results and the statistical analysis. Section 6 presents the discussion and the conclusion. The flowchart of the research method is shown in Fig. 1.

2 Methodology

2.1 Cubic spline interpolation

Spline interpolation is a commonly used method to obtain smooth curves in data processing, and cubic spline interpolation is a more widely used one. Cubic spline interpolation is composed of piecewise cubic curves and has a continuous second derivative at the connecting point, which can ensure the smoothness at the connecting point. Therefore, cubic spline interpolation has the best effect of piecewise low-order interpolation. The cubic spline function is defined as follows:

Usually, on the interval $[A, B]$, $n + 1$ nodes and a set of corresponding function values are given, and if the function satisfies the following:

- 1) $S(x_i) = f(x_i) (i = 0, 1, \dots, n - 1)$ is satisfied at each node,
- 2) There is a continuous second derivative on $n + 1$.

- 3) At every subinterval, $[x_i, x_{i+1}] (i = 0, 1, \dots, n - 1)$ is a cubic polynomial. Then, $S(x_i)$ is called the cubic spline interpolation function.

Cubic spline interpolation polynomial is defined as follows:

The cubic spline interpolation function $S(x)$ is a piecewise cubic polynomial. If asked to work out the function $S(x)$, then four undetermined parameters should be determined between subinterval $[x_i, x_{i+1}]$, if $S_i(x)$ is used to represent its expression on the i th subinterval $[x_i, x_{i+1}]$, then

$$S_i(x) = a_{i0} + a_{i1}x + a_{i2}x^2 + a_{i3}x^3 (i = 0, 1, \dots, n-1) \quad (1)$$

2.2 C5.0 algorithm based on information gain rate

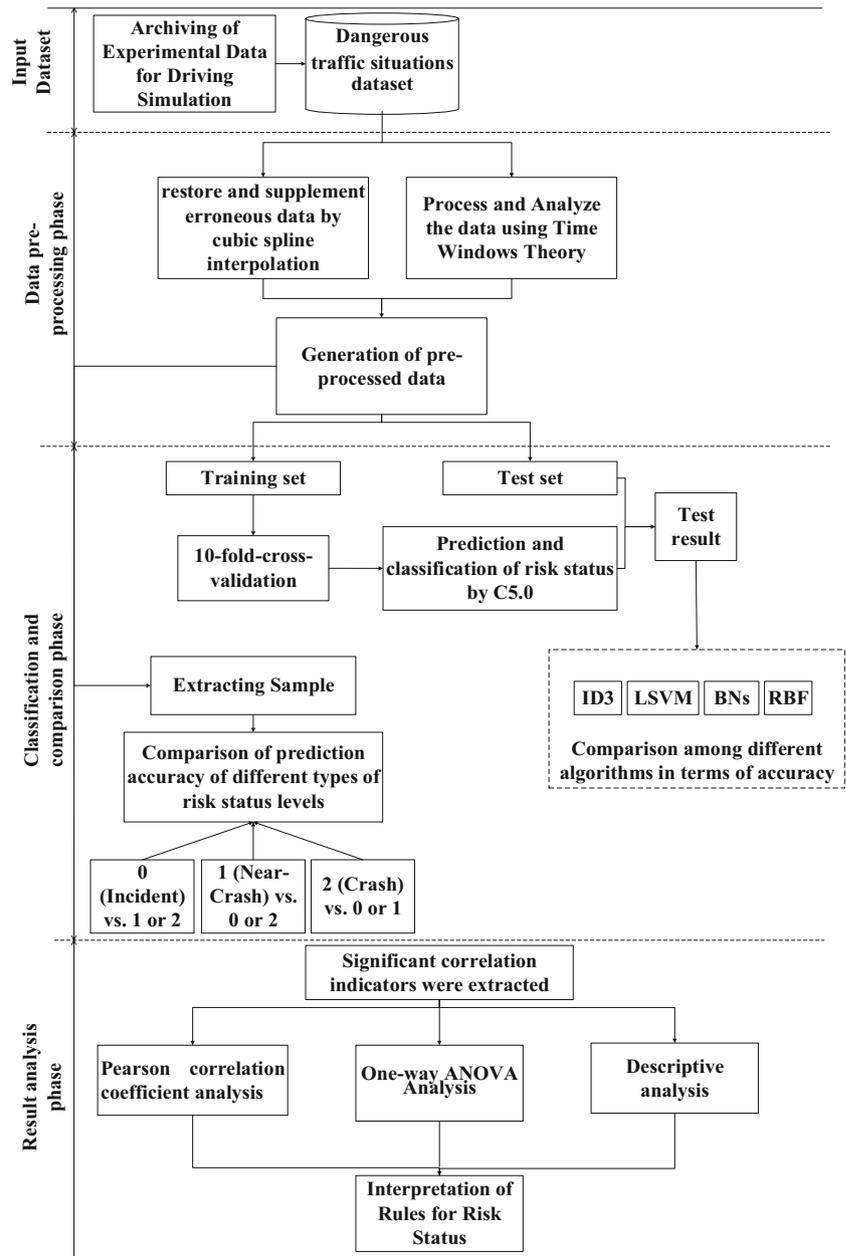
C5.0 algorithm is developed by Quinlan [33], which is an inductive learning method based on the entropy of information in the sample data. Entropy is a measure of the complexity (degree of uncertainty) of a sample set. C5.0 takes the information gain rate as the standard for selecting features, and information gain is the degree of uncertainty reduction. Its decision-making method from the root node calculates the information gain rate of all feature attributes that have the ability to classify the results of the decision tree, and selects the feature attribute pair with the largest information gain rate as the decision tree node. Then, the data are divided into two or more child datasets, and the splitting is stopped when most of the data points are labeled to the similar class within the same branch of the tree. In general, the decision tree model is composed of one root node, multiple intermediate nodes, and several leaf nodes. The root nodes and the intermediate nodes represent the corresponding test conditions, and the leaf nodes represent the final classification results. The main advantage of using the C5.0 algorithm is that the tree structure is more concise and has better storage capacity. In addition, it is faster than the C4.5 decision tree classification and has a better noise reduction effect [34].

The basic principle of C5.0 is to split the sample on the basis of the field that provides the maximum information gain. Then, according to the field division, the subsamples are defined by the first integral of each sample, and the process repeats until the subsamples cannot be divided further. Finally, the bottom segment is checked again, and the segment that has no significant effect on the model values is removed or trimmed.

Let the training set have m samples. Here, m is the number of independent types of $C_i, i = 1, 2, \dots, m$. R_j is a subset of C_i in dataset S, R_i is used to represent the number of tuples in R_j , and the expected value of set S in the classification can be expressed as follows:

$$I(r_1, r_2, \dots, r_m) = - \sum_{i=1}^m p_i \log_2(p_i) \quad (2)$$

Fig. 1 Flowchart of the proposed research method



In the above formula, p_i represents the probability that any sample belongs to C_i , $p_i = \frac{r_i}{|s|}$, and $|s|$ is the tuple in the training set. Let A represent the attribute with a total of v different values $\{a_1, a_2, \dots, a_v\}$. Then, divide the sample set into v subsets according to A . Let S_j be the attribute in dataset S ; the value of S_j is equal to the subset of a_j . In the classification, if A is a decision attribute, the sample set can be divided into different branches. If S_{ij} is used to represent the data of the tuples belonging to class C in the subset, then the entropy of A for C_i , $i = 1, 2, \dots, m$ can be calculated as follows:

$$E(A) = \sum_{j=1}^v \frac{S1j + \dots + Smj}{|s|} I(S1j + \dots + Smj) \quad (3)$$

$$W_j = \frac{S1j + \dots + Smj}{|S|} \quad (4)$$

In the above formula, W_j is the proportion of S_j in S , which can be used as the weight of S_j . The expected $I(S1j, S2j \dots Smj)$ of each value of A for C_i can be calculated as follows:

$$I(S1j + \dots + Smj) = - \sum_{i=1}^m p_{ij} \log_2(p_{ij}) \quad (5)$$

At this moment, $P_{ij} = \frac{S_{ij}}{|S_j|}$ represents the proportion that belongs to C_i in S_j .

With the above calculation and A as the measure of the decision classification attributes, the information gain can be calculated as follows:

$$Grain(A) = I(r_1, r_2, \dots, r_m) - E(A) \tag{6}$$

Because the information gain divides the sample into a smaller subset, there is a certain deviation in the value of the variable. In order to reduce this deviation, the calculation can be expressed as follows:

$$Split\ Info(S, v) = \sum_{i=1}^m \frac{|S_i|}{|S|} \times \log_2 \frac{|S_i|}{|S|} \tag{7}$$

The gain rate is expressed as follows:

$$GrainRatio = \frac{Grain(S, v)}{Split\ Info(S, v)} \tag{8}$$

Table 1 is the decision tree learning algorithm:

Table 1 C5.0 decision tree process

Algorithm: C5.0 decision tree
 Input: Training set $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$;
 Attribute set $A = \{a_1, a_2, \dots, a_n\}$.
 Start: Function Tree Produce(T, A)

- 1: Produce first node;
- 2: If all the samples in T belong to the same category C then
- 3: Mark node as a C -type leaf node; return
- 4: End if
- 5: If $A = \emptyset$ or Sample T is the same as the value in A then
- 6: Mark the node as a leaf node, and mark its category as the category with the largest number of samples in T ; return
- 7: End if
- 8: Select the optimal partition attribute a_* from A ;
- 9: For a_* each value. a_*^v do
- 10: Generate a branch for node, $T_v \subseteq a_*^v$
- 12: Mark the node as a leaf node, and mark its category as the category with the largest number of samples in T ; return
- 13: Else
- 14: Mark TreeProduce($T_v, \frac{A}{\{a_*^v\}}$) as a node for the branch
- 15: End if
- 16: End for

Output: A decision tree with “node” as the root node
 Three conditions for ending the program as follows:
 1) Step 2: All samples in sample T belong to a category, and the division stops.
 2) Step 5: If the attribute set is empty or all the samples have the same attribute value at this time, it is no longer possible to divide according to the attribute, so find the category with the most number among the remaining samples.
 3) Step 12: If the data set does not have samples on an attribute, for example, after multiple partitions, there are no samples of this attribute in the remaining samples, we make the category of such samples equal to the category with the largest number of samples in the current node’s parent node.

2.3 Evaluation methods

The ROC curve (receiver operating characteristic curve, sensitivity curve) analysis method originated from the theory of electronic signal observation and is used to evaluate the radar signal receiving capability. This method is based on the statistical decision-making theory and is currently more mature in applications such as medical diagnosis, human perception and decision-making, military monitoring, and industrial quality control. In this study, we attempted to use this method to distinguish among the different levels of the driving risk status. The ordinate was the true positive rate (*sensitivity*), and the abscissa was a false positive rate ($1 - \textit{specificity}$).

AUC (area under the curve) is an evaluation index used to measure the advantages and disadvantages of the binary classification model. It is generally defined as the area surrounded by the coordinate axis under the ROC curve and ranges from 0.5 to 1.0. Because the ROC curve is generally above the straight line ($y = x$), a larger AUC represents a better performance. The closer the area of AUC is to 1.0, the more rational is the model.

True positives (TP) is the number of positive examples that are actually positive and are classified as positive by the classifier. False positives (FP) is the number of incorrectly classified as positive examples, i.e., the number of instances that are actually negative but are classified as positive by the classifier. False negatives (FN) is the number of instances that are wrongly classified as negative examples, i.e., the number of instances that are actually positive examples but are classified as negative examples by the classifier. True negatives (TN) is the number of instances correctly classified as negative examples, i.e., the number of instances that are actually negative examples and classified as negative examples by the classifier. The true positive rate (TPR) is defined as the percentage of samples correctly judged to be positive in all the samples that are actually positive. The false positive rate (FPR) is defined as the percentage of samples that are misjudged to be positive in all the samples that are actually negative. The ROC curve is a curve that uses the composition method to describe the relationship between the sensitivity (true positive rate, TPR) and the specificity (true negative rate, TNR) of a diagnostic test and reflects the diagnostic test as a comprehensive index of the two continuous variables, sensitivity and specificity. In addition, two other indicators are commonly used in the process of diagnostic tests, namely, the false negative rate (false negative rate, FNR) and the false positive rate (false positive rate, FPR) [35]. These indicators are shown in Eqs. (9)–(12).

The five detailed evaluation indicators are described as follows: Accuracy indicates the proportion of positive samples that are actually positive. In general, in order to distinguish whether a model is good or bad, it is necessary to combine the recall rate (*recall*) and the accuracy rate (*precision*). The recall rate represents the percentage at which all instances

that are actually positive examples are predicted to be positive examples and is equivalent to sensitivity. F-measure is a comprehensive evaluation index that can comprehensively evaluate the accuracy and the recall rate. The higher the F-measure is, the more effective is the model. Accuracy is a common indicator that represents the percentage of correct predictions in all the samples. The indicators are obtained using Eqs. (13)–(16).

$$TPR = \frac{TP}{TP + FN} \quad (9)$$

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

$$TNR = \frac{TN}{FP + TN} \quad (11)$$

$$FNR = \frac{FN}{TP + FN} \quad (12)$$

$$precision = \frac{TP}{FP + TP} \quad (13)$$

$$recall = \frac{TP}{FN + TP} \quad (14)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

$$F\text{-measure} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \quad (16)$$

3 Experiments and results

3.1 Experimental conditions

In order to ensure the safety and the reliability of the experiment, the driving simulation system (shown in Fig. 2), which was developed by ITS Center at Wuhan University of Technology, China, was applied. This system consists of four subsystems: vehicle information data collection system, visual system, sound support system, and the driver's physiological data collection system. The data such as brake signal, vehicle speed, and course angle were collected by this system. In addition, the traffic accident and illegal behaviors were automatically recorded in the driving simulation system.

Two types of data were collected through the data acquisition system. One was the data related to the driver, including blood volume plus (BVP), skin conductance (SC), respiration rate (RR), and PERCLOS, which were mainly used to evaluate the physiological status of the driver. The other was vehicle-related data, and the evaluation of the vehicle status often depended on the vehicle-related data, including speed, acceleration, and steering wheel angle.

3.2 Participants

Thirty-two volunteers recruited on the university campus participated in this study. Two of them dropped out because of dizziness and other symptoms during the experiment, and 30 (22 male and 8 female) volunteers were finally considered. Their ages ranged from 21 to 40 years ($M = 24.6$ years; $SD = 4.8$ years). All of the invited volunteers had valid driving licenses with an average driving experience of 4 years ($SD = 1.5$ years) and were required to operate as though they were driving on an actual road. A summary of the participants' characteristics is given in Table 2.

3.3 Road scenarios

In order to simulate the actual driving process more realistically, we enhanced the fidelity of the simulated driving and the immersion of the subjects. The experimental scene was designed according to the real road data from Wuhan, implemented by using the RoadBuilder software, and loaded into the driving simulator. The parameters of the final scene considered in this experiment are listed in Table 2.

Certain relevant studies have shown that the possibility of vehicle–vehicle and vehicle–human conflict was considerably increased when the traffic flow was large and the vehicles were at the intersections or the bus stations. In order to relatively balance the number of traffic incidents with different driving behavior risk levels considered in this experiment, the number of intersections and bus stations was increased in the scene design process. The actual effect of the scenario is shown in Fig. 3a–d.

3.4 Data collection and driving protocol

During the experimental process, two assistants were recruited; one of the assistants was responsible for debugging the driving simulator and assisting the driver in wearing physiological devices and the other equipment, and the other



Fig. 2 Experimental system

Table 2 Experimental scenario design parameters

Scenario requirements	Detailed design parameters
Time	Day (6:00–18:00)
Weather condition	Sunny; cloudy; sufficiently dry road
Road type	Urban road
Road condition	The road section is 30 km long; the main trunk road is a six-lane two-way traffic road; the secondary trunk road is a four-lane two-way traffic road and includes eight intersections and five bus stations; there are pedestrians walking along the intersection crossings and in some other areas. The landscape greenbelt segregates the non-motor vehicle lanes from the motor vehicle lanes.
Roadside environment	Roadside environment including commercial buildings, billboards, and landscape green belts
Traffic background	Randomly generated bus traffic flow. The main trunk road traffic volume is 2000 vehicles/h; the secondary trunk road traffic volume is 1000 vehicles/h.

assistant functioned as a recorder, recording the driver's individual characteristic information, including the record of special traffic events during the experiment and the record of the self-reported results of the driver. After the experiment, the assistants were responsible for sorting out the data and the video collected by the experiment. The detailed experimental process was as follows:

- Step 1:** Preparation for the experiment: The assistants gave a brief introduction to the experimental procedure and explained the task and the requirements to the volunteers. They informed the volunteers of the rules related to the rewards and the punishments (Each volunteer was paid RMB 100 for completing all of the experimental tasks. If the experiment was terminated early or the requirements were not met, the experiment had to be restarted or no payment was provided). The volunteers were asked to complete individual information questionnaires and sign an informed consent letter. The participants were asked to complete a questionnaire, which included some of the driver's basic personal information (such as the driver's age and gender).
- Step 2:** Introduction: The volunteers were provided the operating specifications and matters that required the attention of the driving simulation system. The staff assisted them in wearing the physiological devices correctly. Then, 10–15 min of adaptive driving was conducted to ensure that the participants could operate the driving simulator efficiently and make the experiment more reasonable. When the adaptive driving was completed, the volunteers rested for 5 min to adjust their physiological and psychological states, and the staff launched the system into the pre-designed driving scenarios, confirming that all of the data output were in the normal state.
- Step 3:** Formal test: Firstly, the participants were asked to complete the experiment within 20 min. During the

driving process, they were asked to obey the traffic rules and to avoid traffic violations, such as retrograding and going out of the driveway. The assistants were asked to record the participants' self-reporting and assess the current risk level of the driving. At the same time, the time of start was recorded. In order to ensure the integrity of the data acquisition, all the volunteers were not allowed to temporarily interrupt the experiment while driving the simulation unless prior permission was obtained from the staff.

- Step 4:** After experiment completion: An assistant held a simple interview with the driver and gave him/her RMB 100 for completing the experiment. The assistant then copied and stored the experimental data.

3.5 Data preprocessing

Because of the instability and the complexity of the entire simulation system, it was inevitable to obtain some discrete incorrect data. Thus, cubic spline interpolation was applied to restore and supplement the artifacts or erroneous data. In all, 150 samples of valid data were selected as the test dataset, and 10 points were randomly selected from this dataset and deleted. Then, cubic spline interpolation was used to restore these points; the corresponding results are shown in Fig. 4. The absolute value of the relative error between the original points and the spline points was calculated as shown in Fig. 5. From Figs. 4 and 5, we inferred that the cubic spline interpolation could restore the reduction of data points effectively.

Considering that a dangerous traffic event is a process event rather than a point event, the time window method should be used when extracting the event. Studies have shown that the length of the time window of traffic incidents has a significant effect on the recognition accuracy of the constructed model. If the time window length is too large, the data will

Fig. 3 Road driving scenarios constructed on the basis of the real road data from Wuhan, China. **a** Pedestrians crossing the road. **b** Bus parking. **c** Roadside buildings. **d** Intersection point



be cumbersome and cannot be accurately calibrated. If the time window is too small, the data characteristics may not be obvious, and accurate decision-making on traffic incidents may not be realized. Therefore, the recognition rate of the target traffic event under different time windows was used as the judgment condition of the time window length selection, and the recognition rate of the target traffic event under different time windows was obtained as shown in Fig. 6.

In the process of selecting the length of the time window, we first selected 0.5 s as the time interval for the calculation. When the cell was determined (that is, the optimal time window was within a certain small interval, the interval was 1 s), 0.1 s was chosen as the time interval. We refined the calibration again and finally obtained the optimal time window length. Figure 5 shows that when the time window length was 3 s, the highest traffic incident recognition rate reached 94%. Therefore, we used 3 s as the time window to process and analyze the data collected in the simulation experiment.

3.6 Parameter settings

To investigate the vehicle's motion characteristics, driving behavior characteristics, and road conditions in different traffic scenarios, a data collection system was developed for the driving simulator, and the vehicle's motion characteristics data were collected in this system. The Biography Infiniti System was equipped to collect the data of the driver's physiological indexes, such as blood volume pulse, skin conductivity, and respiration. An eye movement measurement

instrument was used to collect the data of the characteristics of the eye movement. The data sampling rate, data type, and symbols are presented in Table 3.

4 Prediction model results

4.1 Results of C5.0 classifiers

The C5.0 algorithm was used to classify the different levels of the risk status, and the risk levels were predefined as the

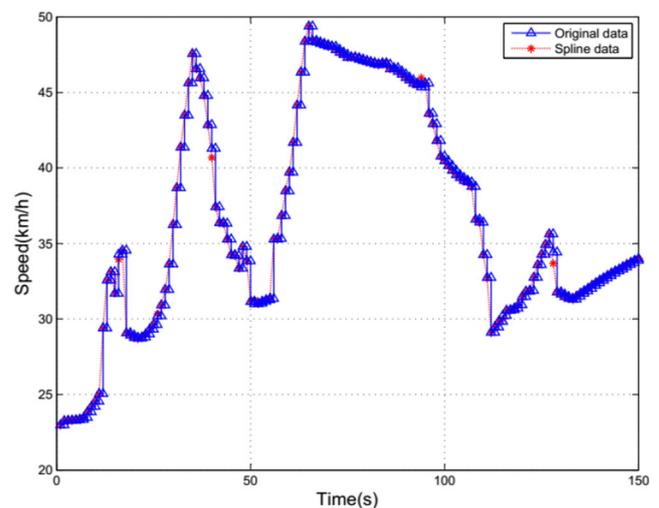


Fig. 4 Result of cubic spline interpolation

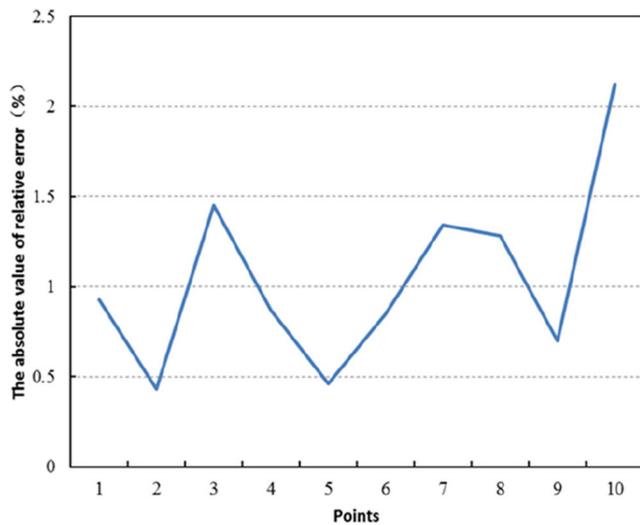


Fig. 5 Error graph

following three levels: 2 - Crash, 1 - Near crash, and 0 - Incident. Moreover, the decision tree was built as follows:

We calculated the information gain rates for all the feature attributes and created the decision tree nodes. Then, all the training sample data were entered as the initial training dataset. The information gain rate of the feature attributes was calculated, and the feature attribute with the maximum information gain rate was selected as the decision tree node. The RESP and SPE attributes with the highest information gain rate were the attributes in the first and the second layers of the decision tree. Starting from the third layer, the classification probability values of FWA, SM, ACC, and the other factors gradually became similar. The risk status decision tree model was established and is illustrated in Fig. 7 (as the complete decision tree is too large, only a part of it is shown).

As shown in Fig. 7a, the C5.0 decision tree method was adopted for the risk status classification. For the purpose of

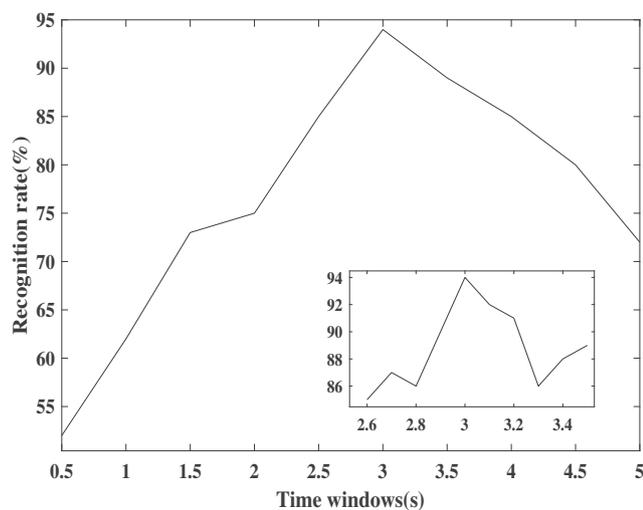


Fig. 6 Time window calibration

explanation and prediction, the result of classification through visual processing was reflected. The first node at the top of the decision tree shows the first optimal split of the risk status, sending cases (respiratory rate) with less than or equal to 39.170 to the left and all the others to the right. In other words, the best variable among all the variables to explain the variability of the classified risk status was RESP. We assumed that when the risk status was at the node, the RESP was greater than 39.170. Under these conditions, the next best classification variable was the vehicle speed (SPE). When SPE was less than or equal to 24 km/h, the risk status path moved to the left, and when the SPE was larger than 24 km/h, the risk status path moved to the right, forming a terminal node or a leaf node. The remaining splits, for the risk status path with SPE less than or equal to 24 km/h, were made on the basis of the front wheel angle (FWA), acceleration (ACC), and SM_FATIGUE (SM). In general, to estimate the path of the risk status, we moved down the branches of the tree in the abovementioned manner until we reached the terminal node. Obviously, as shown in Fig. 7a, the more important variable seemed to be the FWA at the lower risk status, while for the higher risk status, the vehicle speed seemed to be the most important. It seemed reasonable that risky driving was associated with a higher speed. However, when the vehicle was moving at a high speed, the main cause of the accident was driver fatigue, rendering SM as the more important variable.

Figure 7 b shows the result of applying the C5.0 method to classify the three levels of the risk status, which generated a more compact decision tree. The node of the first optimal split was also the variable RESP. Thus, it seemed that RESP was the best variable to classify the risk status into the three levels. Note that FWA was an important variable in the case of a low vehicle speed, and for a higher speed, the driver’s fatigue was more important in the safe driving process. Thus, respiration (RESP) and SM_FATIGUE (SM) once again proved to be important variables when the level of risk increased. For the purpose of prediction, the two-level risk status path was used to check the three-level risk status, as their tree structure was similar, while the decision tree was used to determine the prediction. For instance, suppose that we had to make a risk status prediction for the sample. First of all, from the top of the decision tree to the root node, we branched right (RESP >39.170), left (SPE ≤24.000), right (FWA >1.771), or right (SM >0.044) to obtain the risk status level of 0. Some variables were selected several times during the algorithm classification process. Considering that all the left branches were to the leaf node, the SPE appeared twice. Because one of the goals of C5.0 was to categorize the data and generate a tree-like structure, relatively few variables appeared explicitly in the segmentation criteria, and some very important variables appeared more than once (such as the SPE in this tree structure). This might mean that these variables were not important to the dependent variable at the time of prediction but could be considered very

Table 3 Independent variable description

ID	Variables	Symbol	Type	Sampling rate (Hz)
1	Speed	SPE	Continuous	20
2	Brake signal	BRA	Continuous	20
3	Left-turn	LT	Qualitative	20
4	Right-turn	RT	Qualitative	20
5	Front wheel angle	FWA	Qualitative	20
6	Acceleration of front wheel angle	AFWA	Qualitative	20
7	Acceleration	ACC	Continuous	20
8	Course angle	CA	/	20
9	Pitch angle	PA	/	20
10	Tilt angle	TA	/	20
11	Blood Volume pulse	BVP	/	256
12	Skin conductivity	SC	/	20
13	Respiration	RESP	/	32
14	The distance to center lane	CLD	/	15
15	The distance to the left lane	LLD	/	15
16	The distance to the right lane	RLD	/	15
17	SM_FATIGUE	SM	/	32
18	MICROSLEEP	MIC	/	32
19	PERCLOS	PER	/	50
20	Acceleration(X)	ACCX	/	50
21	Acceleration(Y)	ACCY	/	50
22	Acceleration(Z)	ACCZ	/	50
23	Risk Status	RS	Qualitative	/

important to the independent variable, even if it never appeared as the primary segmentation node.

4.2 Results of model test

Based on the simulation experiment, a large amount of sample data was acquired and selected; we used 100 groups of the sample data to test the C5.0 prediction accuracy of the algorithm. The comparison between the results of the true value and the predicted value is shown in Fig. 8, and Table 4 shows the test accuracy. As the chart shows, the prediction had high accuracy. When predicting the three-level risk status, only six test samples incorrectly predicted the risk status. In particular, the prediction results for the risk status levels of 2 and 1 only had three and two sample prediction errors, respectively.

As can be seen from the test results given in Table 4, in the sample test, the recognition accuracy of C5.0 in the third-level risk state was almost 94%. Figure 8 uses C5.0 to establish a mapping relation of the sample comparison at different risk states, which indicated that the real value corresponded to the predicted value. For example, if the true value was 2 and the corresponding prediction value was 1, the prediction of the individual sample was inaccurate. The higher the number of samples was, the lower was the prediction accuracy (Table 4).

The ROC curves for the three-level risk status were generated to evaluate the performance, and the corresponding results are shown in Fig. 9. The risk status classification based on the C5.0 algorithm demonstrated high predictive power for the three different risk status levels [0 vs. 1 or 2 (Fig. 9a), 1 vs. 0 or 2 (Fig. 9b), and 2 vs. 1 or 0 (Fig. 9c)]. The areas under the curve (AUC) (Fig. 9d) reached 0.934, 0.77, and 0.845. Moreover, the ROC curves showed that the risk level of 0 was the most accurate one (the AUC reached 93.40%, which was close to the ideal value of 1, and the accuracy was 76.8%) when identifying the different levels of the risk status. The 10-fold cross validation method was used to evaluate the accuracy of the C5.0 algorithms.

4.3 Results of different algorithms

Figure 10 lists six evaluation indexes used to evaluate the classification performance of these five algorithms, which is C5.0 decision tree (C5.0), Radial Basis Function (RBF), Lagrangian Support Vector Machines (LSVM), Iterative Dichotomiser 3 (ID3), and Bayesian network (BNs), respectively. The results showed that these five algorithms could classify the risk states, but the classification effect was different. In addition, the best FPR (0.145) was obtained by the ID3

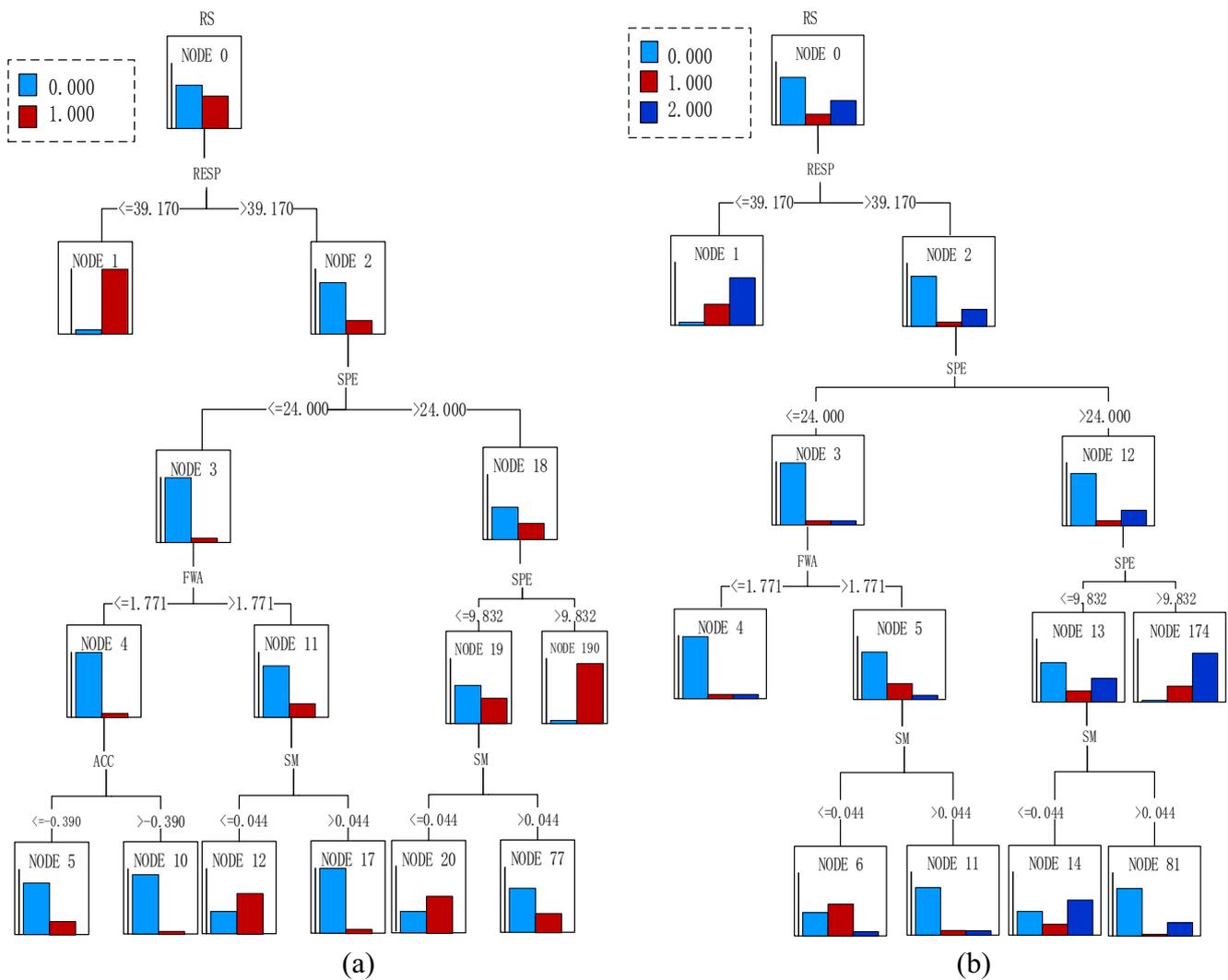


Fig. 7 Results of decision trees. **a** Decision tree for level-two risk status. **b** Decision tree for level-three risk status

Fig. 8 Prediction results of test set by C5.0

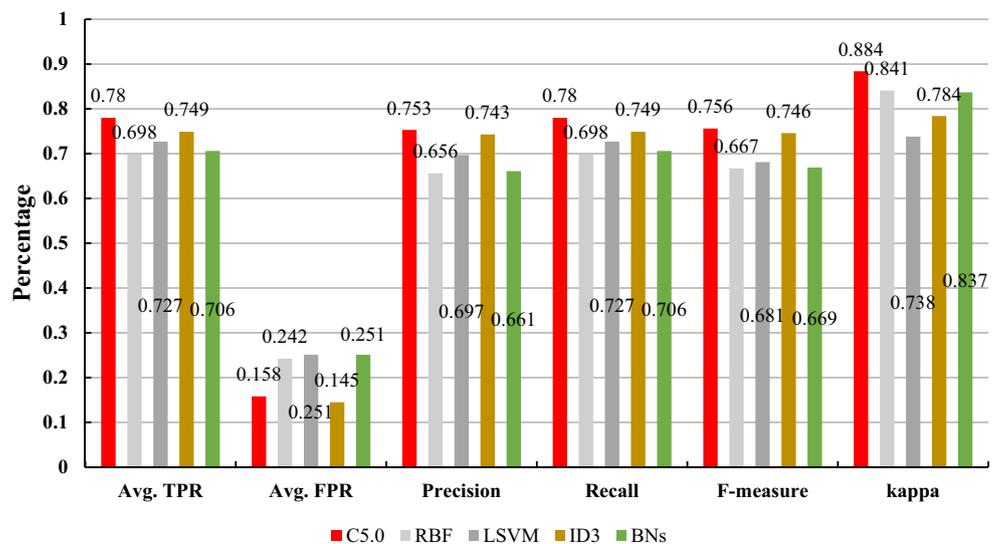


Table 4 Identification accuracy rate based on C5.0 (test sample)

Driving risk level	Accuracy amount	Accuracy rate	Missing amount	Missing rate
0	54/55	98.10%	1/55(1)	1.81%
1	9/12	75%	2/12(0)1/12(1)	25%
2	31/33	93.93%	1/33(0)1/33(1)	6.06%
Average	94/100	94%	6/100	6%
0	55/55	100%	0/55	0%
1	23/45	51.11%	22/45	49%
Average	78	78%	22	22%

algorithm (The lower the value of FPR, the better was the effect of the prediction.). The C5.0 achieved the best TPR (0.78), precision (0.753), recall (0.78), and kappa (0.884), which proved that C5.0 performed better than the other four models. Although the FPR in C5.0 was not the best among the five algorithms, it did reach 0.158 and was second only to that of ID3, which was 0.145.

Figure 11 shows the performance comparison of the five algorithms. In general, the higher the ratios of TPR to FPR and the other four indicators (precision, recall rate, F-measure, and kappa) were, the better was the classification method. In this

study, six performance evaluation indexes of the algorithms were selected for comparison. The kappa statistic of the algorithms (C5.0 > RBF > BNs > ID3 > LSVM) indicated that among the five algorithms, the difference between the predicted value and the actual value was minimal when using C5.0, which implied that the C5.0 algorithm had the highest accuracy. In terms of the FPR, the C5.0 had worse performance than the ID3 algorithm but better performance than the RBF, LSVM, and BNs. Overall, the C5.0 was optimal among the five algorithms and was effective enough to be used to classify the different levels of risk status.

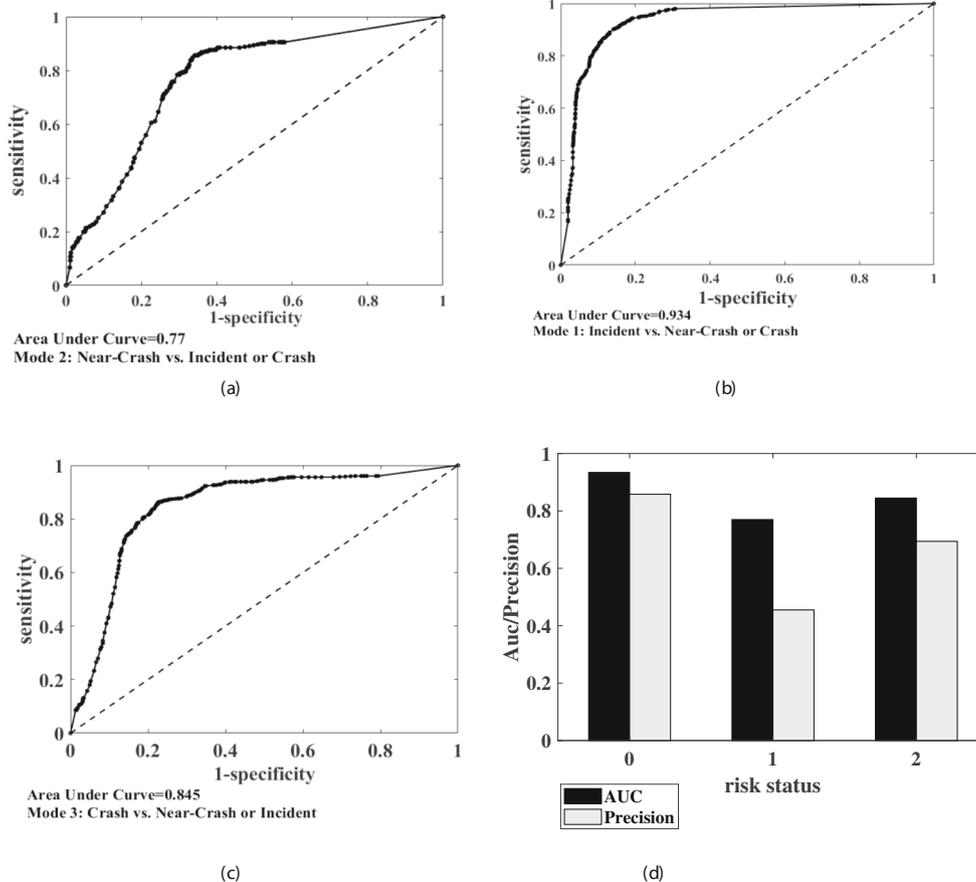


Fig. 9 ROC curve. **a** 0 vs. 1/2, **b** 1 vs. 0/2, **c** 2 vs. 1/0, and **d** area under curve/precision

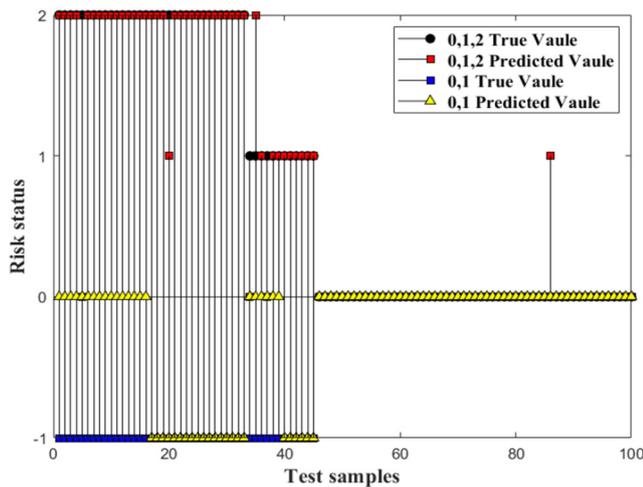


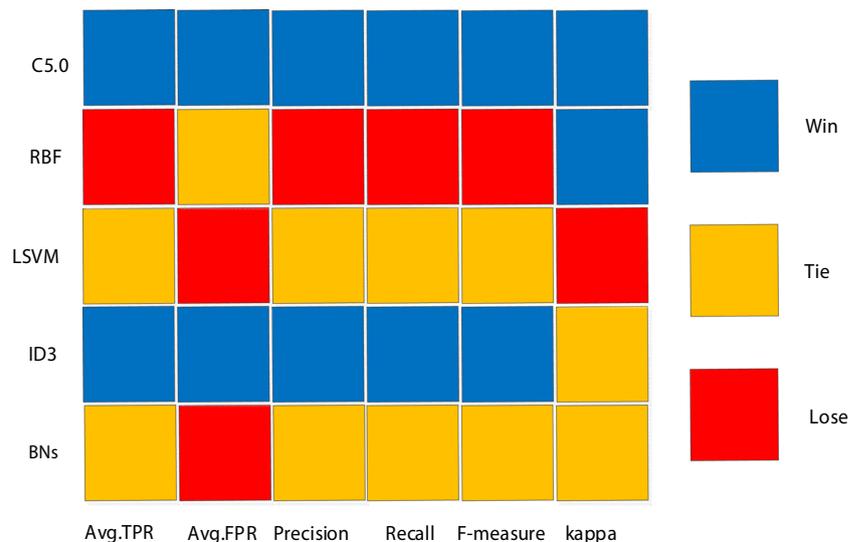
Fig. 10 Accuracy of the five algorithms

5 Result analysis

5.1 Statistical analysis

Pearson’s correlation analysis was used to examine the correlation between the eight influencing factors and the risk status. As shown in Table 5, the results showed that six factors (RESP, SPE, SM, LLD, CA, and SC) were significantly correlated with the risk situations ($p < 0.05$). Three factors (FWA, ACCX, PER) were weakly correlated with risk status. Among the significantly correlated factors, three factors were negatively correlated with the risk status, namely, RESP, SM, and LLD, which indicated that with the increase of the risk status, all these four factors presented a downward trend. The remaining three significant correlation factors (SPE, CA, SC) showed a positive correlation with the risk status, indicating that these six factors also showed an increasing trend with the increase of the risk status. We all know that the faster, the more dangerous. Therefore, it is not difficult to understand

Fig. 11 Performance comparison of the five classification algorithm



that an increase in speed, skin conductance, and course angle leads to an increase in increased risk status. It can also be known from the table that the SPE is positively correlated with CA and SC respectively, which indicates from another perspective that the three increase when the risk status rises at the same time. Similarly, it seems reasonable that SM_FATIGUE, the distance to the left lane, and respiration will decrease as the risk status increases.

5.2 One-way ANOVA analysis

As can be seen from Table 6, all the eight factors except ACC have significant differences. From the correlation analysis, ACC is also weakly correlated with risk status, and it can be seen from the analysis of the two that the rise of risk status is not closely related to ACC. LSD (least significant difference) showed that SPE, SC, FWA, ACCX, and CA of level 0 risk status are significantly less than levels 1 and 2, and RESP and SM are significantly greater than levels 1 and 2. The analysis of LSD showed that the respiration decreased from the beginning of the accident to the end of the crash. SM_FATIGUE results show that when accidents occur, they are usually caused by a high degree of fatigue. The purpose of this study was to classify and predict the risk status of current driving based on an improved decision tree (C5.0) algorithm through the driving performance indicators. The on-road experiment collected multi-sensor data, and the relationship between dangerous driving behaviors and influencing factors was visualized through decision trees. The repeated measurement analysis of variance (ANOVA) was used to analyze the specific influence characteristics of the physiological indicators and the vehicle indicators of 30 drivers under the risk state, and the results of ANOVA are shown in Table 6. The results confirmed that these six factors had a strong relationship with the risk status. From the two abovementioned analyses, we

Table 5 Correlation of the factors

Cor.	RESP	FWA	SPE	SM	LLD	ACC	CA	ACCX	PER	SC	RS
RESP	1	0.028	0.475**	0.079**	0.026	0.032	0.082**	0.070**	0.122**	0.019	0.549**
FWA		1	0.070**	0.033	0.000	0.015	0.059**	0.041*	0.007	0.008	0.049*
SPE			1	0.162**	0.049*	0.015	0.165**	0.066**	0.003	0.130**	0.489**
SM				1	0.012	0.027	0.029	0.056**	0.081**	0.102**	0.121**
LLD					1	0.050**	0.021	0.035	0.013	0.032	0.061**
ACC						1	0.076**	0.078**	0.007	0.056**	0.033
CA							1	0.004	0.011	0.098**	0.081**
ACCX								1	0.033	0.006	0.049*
PER									1	0.040*	0.038*
SC										1	0.052**
RS											1

* Correlation is significant at the 0.05 level; ** Correlation is significant at the 0.01 level

inferred that all the six driving factors (respiration, left lane distance, speed, SM_FATIGUE, course angle, and skin conductivity) were significantly affected by the driving risk status. Pearson’s correlation analysis of the results between the two indicators clearly showed the source of the significant difference.

5.3 Factor analysis

Six strongly correlated indicators were selected for the descriptive analysis. The statistical deviation of the respiration for the three levels of risk status was 0.943, 9.596, and 9.817, respectively, as shown in Fig. 12a, which indicated that an increase in the level of the risk status could cause a more significant change in respiration while driving. The means and the standard deviations of the other five factors are presented in Fig. 12b–f. As can be observed, the mean of the vehicle speed in level 2, which reached approximately 80.81 km/h (SD = 23.19 km/h), was the highest of the three levels. The values of the SM (mean = 0.065, SD= 0.035) and the left lane distance (mean = 1.74 m, SD = 1.32 m) in the case of the risk status of level 0 were the highest. In addition, when the level of risk status was 0, the values of the skin conductivity

(mean = 8.44, SD = 0.928) and the course angle (mean = 146.411, SD = 34.8527) were the lowest. Finally, the one-way ANOVA results ($p \leq 0.001$) shown in Table 6 indicated that the values of the six factors in the three different levels of the driving risk status had significant differences. When the significance level was 0.05, the three levels of the risk status could be predicted by using these six features.

6 Discussions and conclusions

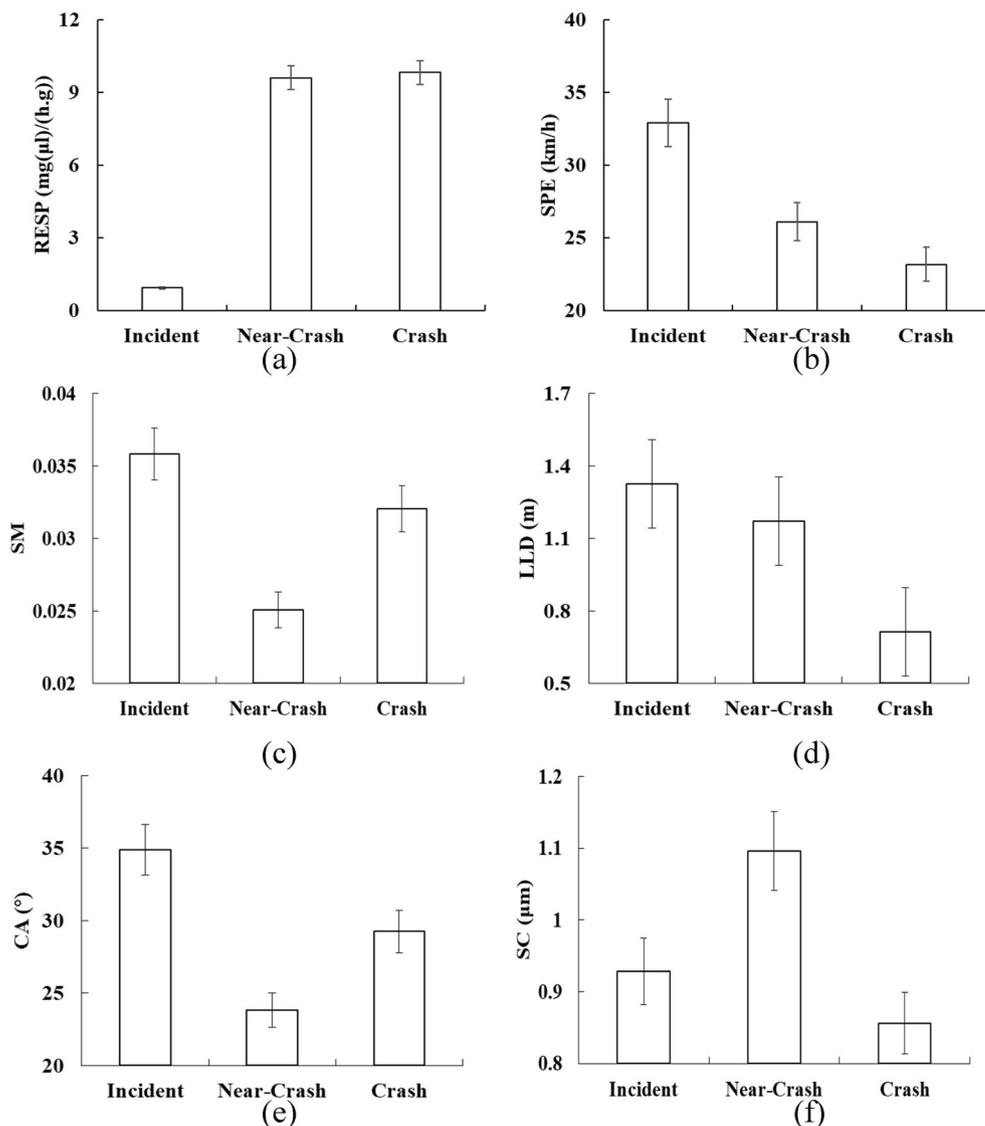
As shown in Fig. 13, the C5.0 algorithm was compared with the LSVM, RBF, ID3, and BN algorithms. A number of performance evaluation indexes were applied to compare the five models. In general, the prediction accuracy of the C5.0 algorithm for the risk status was 78%, which was significantly higher than that of the other algorithms. Thus, a risk status prediction model using C5.0 was the best method to predict risk levels.

Figure 14 shows the accuracy of the corresponding risk status obtained using five different models. C5.0 scored the highest for the three levels of risk status. When the risk level was 0 and 2, respectively, the prediction and the recognition

Table 6 One-way ANOVA analysis

Risk status	SPE Mean	FWA Mean	RESP Mean	SC Mean	LLD Mean	ACC Mean	SM Mean	ACCX Mean	CA Mean
Incident (0)	45.258	0.732	54.985	8.445	1.897	0.136	0.066	0.055	146.411
Near-crash (1)	77.379	1.669	45.481	8.746	1.873	0.156	0.050	0.052	156.145
Crash (2)	80.811	1.485	45.451	8.535	1.739	0.175	0.058	0.052	151.671
Sig.	0	0.015	0	0	0.004	0.232	0	0.015	0
F	469.833	4.189	690.552	15.066	5.441	1.463	39.257	4.195	16.479
LSD	0<1,2	0<1, 2	0>1, 2	0<1, 2	0> 2		0>1, 2	0<1, 2	0<1, 2

Fig. 12 Characteristics analysis for different levels of driving risk status. **a** Respiration. **b** Speed. **c** SM_FATIGUE. **d** Left lane distance. **e** Course angle. **f** Skin conductivity



accuracy of the C5.0 algorithm were 0.858 and 0.694, respectively. When the risk level was 1, C5.0 was second only to LSVM (it achieved the best accuracy of 0.524). All in all, the C5.0 algorithm proved to be the optimal method for identifying different types of risk states.

The risk status experiment involved the collection of various types of driving data, such as vehicle speed, heading angle, front wheel angle, and the driver’s physical indexes. These different types of data made it possible to predict the driving risks in a timely manner while driving. The C5.0 results showed that the driving risk conditions of the different drivers and for the different situations differed significantly, and these changes were consistent with actual road driving. At the same time, they proved that other factors such as human–vehicle–road conditions and the environment also affected the occurrence of dangerous driving conditions. In addition, the

results of Pearson’s correlation, one-way ANOVA, and the classification indicated that respiration, left lane distance,

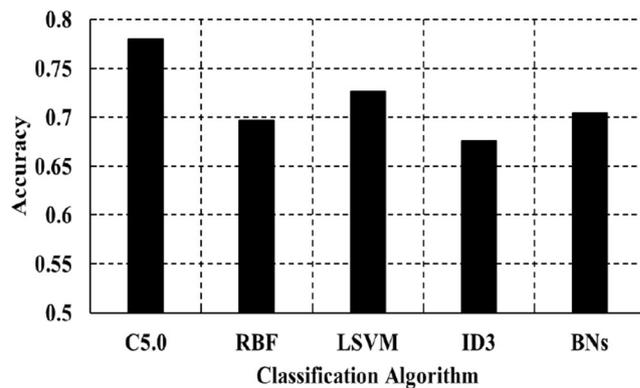


Fig. 13 Average accuracy of different classification algorithms

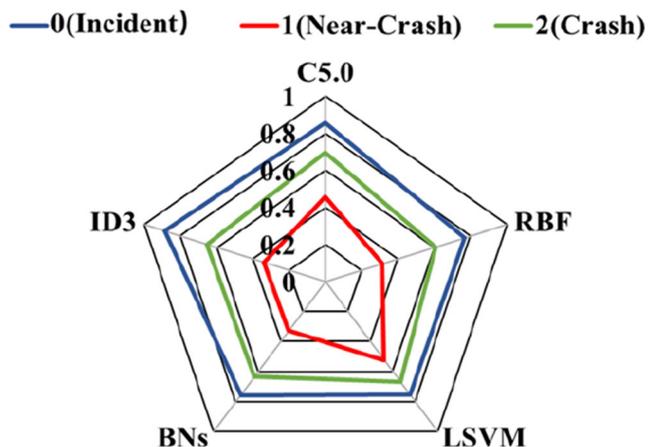


Fig. 14 Radar maps of different classification algorithms

vehicle speed, SM_FATIGUE, course angle, and skin conductivity had significant influences. Therefore, these six factors were taken as the independent variables to establish the driving risk prediction model.

In this study, we mainly investigated whether these six factors could effectively predict the driving risk status. Therefore, on the basis of the simulated experimental data, C5.0 was used to establish the driving risk status prediction model. In order to ensure that the model could effectively and timely identify driving risks, 30 samples of driver simulation experiments were collected and the accuracy of five different prediction models was tested. These results indicated that C5.0 was highly accurate in predicting the driving risk status. The traffic accidents were caused by risky driving, but the frequency of traffic accidents was occasionally high in daily life. In various real driving scenarios, it is difficult to assess different driving risk states, but virtual experiments provide an opportunity to assess the state of driving risk by formulating specific scenarios or events. Through the analysis of the experimental data, we found all the factors leading to a traffic accident. A previous study [36] used natural driving data to predict the risk of individual drivers. However, because of the consideration of the sample size, collisions and near-collision accidents were considered to be collisions in the analysis, which led to the problem of defining a near-collision [37]. In a driving simulator experiment, this problem is easy to solve because there is absolute safety. Most of the accident data can be acquired through simulation experiments, which are a good means to assess the driving risk status. This study had some limitations. First, because of the study's voluntary nature, the sample might be not sufficiently large with only 30 participants. Such a small sample size might ignore the relationship between individual differences in drivers and driving risk factors. Moreover, we only analyzed the simulation data for a specific scenario, and the driving risk status varies in different scenarios and for different types of roads. The scenario reported in this paper is that of the most common type of road, and

whether this model can be extended to other scenarios, such as highways and mountain roads, remains to be studied. In the future, the driving risk status prediction model should include more aspects such as the driver's personality and the type of vehicle.

Funding This study is sponsored by the National Natural Science Foundation of China under Grants 51805169, 52072288. This study is also supported by Natural Science Foundation of Jiangxi Province under Grant 20202BABL212009. This work is also sponsored by the National Natural Science Foundation of China under Grant 52072288. This study is also sponsored by the Special Fund for Graduate Student Innovation of Jiangxi Province (YC2020-S330).

Data availability The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest The authors declare no competing interests.

References

- Zhou T, Zhang J (2019) Analysis of commercial truck drivers' potentially dangerous driving behaviors based on 11-month digital tachograph data and multilevel modeling approach. *Accid Anal Prev* 132:105256–105256
- Dingus TA, Klauer SG, and Neale VL, 2006. The 100-Car naturalistic driving study phase II – results of the 100-Car field experiment, *Chart* (2006) no, HS-810 593.
- Graham R, Carter C (2001) Voice dialling can reduce the interference between concurrent tasks of driving and phoning. *Int J Veh Des* 26:30–47
- Dula CS, Geller ES (2003) Risky, aggressive, or emotional driving: addressing the need for consistent communication in research. *J Saf Res* 34:559–566
- Kamijo S, Matsushita Y, Ikeuchi K, Sakauchi M (2000) Traffic monitoring and accident detection at intersections. *IEEE Trans Intell Transp Syst* 1:108–118
- Charlton SG (2009) Driving while conversing: cell phones that distract and passengers who react. *Accid Anal Prev* 41:160–173
- Caird JK, Willness CR, Steel P, Scialfa C (2008) A meta-analysis of the effects of cell phones on driver performance. *Accid Anal Prev* 40:1282–1293
- Guo Z (2009) Theories and methods on driving risk status identification. Southwest Jiaotong University (In Chinese), Chengdu
- Sagberg F, Selpi, Piccinini GFB, Engstrom J (2015) A review of research on driving styles and road safety. *Hum Factors* 57:1248–1275
- Chen S-W, Fang C-Y, Tien C-T (2013) Driving behaviour modelling system based on graph construction. *Transport Res Part C-Emerg Technol* 26:314–330
- Z. Chen, Y. Zhang, C. Wu, and B. Ran, (2019) Understanding individualization driving states via latent Dirichlet allocation model, *IEEE Intell Transport Syst Mag* 1-1.
- Simons-Morton BG, Klauer SG, Ouimet MC, Guo F, Albert PS, Lee SE, Ehsani JP, Pradhan AK, Dingus TA (2015) Naturalistic teenage driving study: findings and lessons learned. *J Saf Res* 54: 41–44

13. Macadam CC (2003) Understanding and modeling the human driver. *Veh Syst Dyn* 40(1-3):101–134
14. Dehban A, Sajedin A, and Menhaj MB, (2016) A cognitive based driver's steering behavior modeling, 2016 4th International Conference on Control, Instrumentation, and Automation (ICCIA).
15. Hills PJ, Catherine T, Michael PJ (2018) Detrimental effects of carryover of eye movement behaviour on hazard perception accuracy: effects of driver experience, difficulty of task, and hazardousness of road. *Transport Res Part F Traffic Psychol Behav* 58:906–916
16. Lantieri C, Lamperti R, Simone A, Costa M, Vignalia V, Sangiorgia C, Dondia G (2015) Gateway design assessment in the transition from high to low speed areas. *Transport Res F: Traffic Psychol Behav* 34:41–53
17. Li G, Wang Y, Zhu F, Sui X, Wang N, Qu X, Green P (2019) Drivers' visual scanning behavior at signalized and unsignalized intersections: a naturalistic driving study in China. *J Saf Res* 71: 219–229
18. Jimenez-Pinto J, Torres-Torriti M (2012) Face salient points and eyes tracking for robust drowsiness detection. *Robotica* 30:731–741
19. Zhang W Eye gaze estimation from the elliptical features of one iris. *Opt Eng* 50(2011):047003–047009
20. Abdelwahab H, Abdel-Aty M (2001) Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transport Res Record J Transport Res Board* 1746:6–13
21. Hernandezgress N, and Esteve D, (1995) Multisensory fusion and neural networks methodology: application to the active security in driving behavior, Steps Forw Intell Transport Syst World Congress.
22. Huang C, Wang X, Cao J, Wang S, and Zhang Y, (2020) HCF: a hybrid CNN framework for behavior detection of distracted drivers, *IEEE Access* PP, 1-1.
23. Jeberson W, Kishor A, Chakraborty C (2021) Intelligent healthcare data segregation using fog computing with internet of things and machine learning. *Int J Eng SystModel Simul* 1(1):1
24. Sant A, Garg L, Xuereb P, and Chakraborty C, (2021) A novel green IoT-based pay-as-you-go smart parking system. *Computers, Materials and Contin.*
25. Wang J, Zheng Y, Li X, Yu C, Kodaka K, Li K (2015) Driving risk assessment using near-crash database through data mining of tree-based model. *Accid Anal Prev* 84:54–64
26. Li W, Shimin L, Jingfeng Y, Nanfeng Z, Ji Y, Yong L, Handong Z, Feng Y, Zhifu L (2017) Dynamic traffic congestion simulation and dissipation control based on traffic flow theory model and neural network data calibration algorithm. *Complexity* 2017:1–11
27. AlKheder S, AlRukaibi F, and Aiash A (2020) Risk analysis of traffic accidents' severities: an application of three data mining models, *ISA transactions*.
28. Cura A, Kucuk H, Ergen E, and Oksuzoglu IB, (2020) Driver profiling using Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN) methods, *IEEE Transactions on Intelligent Transportation Systems* PP, 1-11.
29. Kumagai T, Akamatsu M (2006) Prediction of human driving behavior using dynamic Bayesian networks. *LEICE Trans Inform Syst* E89D:857–860
30. Chen Z, Cai H, Zhang Y, Wu C, and Sotelo MA, (2019) A novel sparse representation model for pedestrian abnormal trajectory understanding, *Expert Syst Appl* 138.
31. Yan L, Huang Z, Zhang Y, Zhang L, Zhu D, Ran B (2017) Driving risk status prediction using Bayesian networks and logistic regression. *IET Intell Transp Syst* 11:431–439
32. Panagopoulos G, Pavlidis I (2020) Forecasting markers of habitual driving behaviors associated with crash risk. *IEEE Trans Intell Transp Syst* 21:841–851
33. Quinlan R (2004). *Data mining tools See5 and C5.0*.
34. Breiman L, (1984) *Classification and regression trees*, Wadsworth International Group.
35. Berry M, Lino G (2004) *Data mining techniques*. Wiley, Indianapolis
36. Guo F, Fang Y (2013) Individual driver risk assessment using naturalistic driving data. *Accid Anal Prev* 61:3–9
37. Xu X-y, Liu J, Li H-y, Hu J-Q (2014) Analysis of subway station capacity with the use of queueing theory. *Transport Res Part C-Emerg Technol* 38:28–43

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.